DOCUMENT RESUME

ED 438 311 TM 030 622

AUTHOR Cool, Angela L.
TITLE A Review of Methods for Dealing with Missing Data.
PUB DATE 2000-01-28
NOTE 34p.; Paper presented at the Annual Meeting of the Southwest Educational Research Association (Dallas, TX, January 27-29, 2000).
PUB TYPE Information Analyses (070) -- Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Data Analysis; Estimation (Mathematics); Regression (Statistics); *Research Methodology
IDENTIFIERS *Missing Data

ABSTRACT

Missing data occur in virtually every study. This paper reviews some of the various strategies for addressing this problem. The paper also provides instructional detail on two accessible ways of estimating missing data, both using the Statistical Package for the Social Sciences for Windows: (1) substitution of missing values with the variable mean of nonmissing scores; and (2) replacement of missing values with estimates derived from regression. Nine tables and five appendixes provide details of the analyses and outputs. (Contains 12 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

Running head: DEALING WITH MISSING DATA

ED 438 311

A Review of Methods for Dealing with Missing Data

Angela L. Cool

Texas A&M University 77843-4225

TM030622

Angela Cool

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Paper presented at the annual meeting of the Southwest Educational Research Association, Dallas, January 28, 2000.

Abstract

Missing data occur in virtually every study. The present paper reviews some of the various strategies for addressing the missing data problem. The paper also provides instructional detail on two accessible ways of estimating missing data, both using SPSS for Windows: (a) substitution of missing values with the variable mean of non-missing scores; and (b) replacement of missing values with estimates derived from regression. Nine tables and five appendices provide details of the analyses and outputs.

A Review of Methods for Dealing with Missing Data

Missing data are a common problem in empirical research and occur in essentially every study. Indeed, for virtually any large data set it is unlikely that information will be complete for all the cases. For example, it is not uncommon that some information is either missing or in an unusable form when using attitude and behavior measures (Kim & Curry, 1977). Respondents may not answer every item through inadvertence, because they may consider some questions intrusive, or because some items are perceived to be ambiguous. Data might also be missing for a variety of other reasons, the most common being errors in the implementation of the study, interviewer errors (omitted questions, illegible recording of responses, etc.), inadmissible multiple responses to a given item, the loss of instruments, and attrition in the case of a panel-design sample (Anderson, Basilevsky, & Hum, 1983). In fact, it is not unusual in a large data set for a large minority of participants (e.g., 20% or 30%) to have missing data on one or a few items.

When participants do not provide data on a substantial number of items, it seems clear that such persons should be omitted from further analysis. But it seems unnecessary and undesirable to delete 20% or 30% of a sample only because these participants skipped one or two different from among numerous

items, because these participants did provide so much information on the items they answered.

The question naturally arises: how can small amounts of missing data be estimated utilizing the available data? Numerous solutions to this problem have been proposed, which vary considerably in their complexity (Glasser, 1964; Cohen & Cohen, 1983; Little & Rubin, 1987). Often when researchers are faced with such missing data problems, they are likely to select either a listwise deletion or pairwise deletion method, and then proceed to interpret the resulting statistics as usual (Kim & Curry, 1977).

But as the recently released report of the APA Task Force on Statistical Inference emphasized:

> Special issues arise in modeling when we have missing data. The two popular methods for dealing with missing data that are found in basic statistics packages--listwise and pairwise deletion of missing values--are *among the worst methods available* for practical applications. (Wilkinson & The APA Task Force on Statistical Inference, 1999, p. 598, emphasis added)

The primary objective of the present paper is to review and organize some of the various strategies for addressing the missing data problem. To make the task manageable, the

discussion is confined to four accessible ways of dealing with missing data: listwise deletion, pairwise deletion, mean substitution, and regression estimation. The paper also provides instructional detail on the two imputed means analyses, mean substitution and regression estimation, illustrated in the context of SPSS in order to provide the reader with a more concrete understanding of the methods and procedures.

Listwise Deletion

The most obvious method for dealing with incomplete data is to let the computer program discard all cases with any missing values and then use the remaining records to compute results. For most statistical programs, this occurs by default. However, a serious limitation of this approach is that relevant data are frequently discarded (Kim & Curry, 1977; Raymond & Roberts, 1987). Although listwise deletion is the simplest, as the number of variables increases, increasing amounts of data are ignored, even as the total number of missing values remain constant (Raymond & Roberts, 1987). For example, in an extreme case where all respondents in the sample have only one (not necessarily the same) variable missing, the listwise deletion method would discard all cases.

This common practice of discarding all individuals from whom at least one variable is missing leads to excessive loss of statistical power. As the number of cases decreases, there is a

decrease in error degrees of freedom yielding a loss of statistical power and a larger standard error (Cohen & Cohen, 1983; Witta & Kaiser, 1991).

Pairwise Deletion

Pairwise deletion is an attractive alternative when there are a small number of missing cases on each variable relative to the total sample size, and a large number of variables are involved (Kim & Curry, 1977). With this piecemeal method, all available observations for each particular variable are used to compute means and variances, while all available pairs of values are used to compute covariances (Raymond & Robert, 1987). Thus, correlations are computed using only those observations that have nonmissing values on both variables.

The problem with the use of pairwise deletion is the potential inconsistency of the covariance matrix in a multivariate context. When correlations and other statistics are based on different but overlapping subsamples of a larger sample, the population to which generalization is sought is no longer clear. It is possible to compute correlation matrices with mutually inconsistent correlations.

Mean Substitution

Mean substitution assumes that a missing value for an individual on a given variable is best estimated by the mean (expected value) for the non-missing observations for that

variable (Anderson, Basilevsky, & Hum, 1983). However, as Little and Rubin (1987) noted, there a several limitations of the procedure,

> (a)sample size is overestimated, (b)variance is underestimated, (c)correlations are negatively biased, and (d)the distribution of new values is an incorrect representation of the population values because the shape of the distribution is distorted by adding values equal to the mean.(p. 5)

This attenuation of the correlation coefficient, the reduction of r_{XY} as a result of methodological error, can be explained as follows: Remember that the formula for correlation can be expressed as: $r_{XY} = COV_{XY} / (SD_X * SD_Y)$ (Walsh, 1996). The numerator of this formula can be expressed as:

$$COV_{XY} = (\Sigma (X_i - \bar{X})(Y_i - \bar{Y})) / n-1.$$

When we substitute the mean for a given i-th person's missing X_i or Y_i score, that person's deviation score (x_i) will necessarily be zero. Thus, for any person for whom we estimate missing data, the numerator of the COV--the product of that person's deviation scores--will be 0, making that person push the correlation closer to 0. Obviously, we do not want to use mean substitution with too many cases, especially when the correlation coefficients are already fairly close to zero.

Mean substitution sounds simple enough. For a given item, simply substitute the mean response of all valid cases providing data on that item. However, doing mean substitution in SPSS requires a multi-stage substitution process, which is not difficult, but also is not obvious.

SPSS for Windows: Mean Substitution

Assume that we have a sample of size (n=12) in which data are collected in a case (respondent) by variable (response) matrix **X** with rows denoting individuals and columns denoting variables. However for some individuals, one of the responses is missing. The incomplete data matrix in Table 1 shows that person 1 is missing data on the first variable (X1), person 2 is missing data on the second variable (X2), person 3 is missing data on the third variable (X3), and person 4 is missing data on the fourth variable (X4).

INSERT TABLE 1 ABOUT HERE.

Step One: Input Data and Count Number of Missing Entries.

When entering data into a Word document, be sure to use a fixed width font (e.g., courier) and save the document as 'text only'. Leave blanks for missing observations. This allows the researcher to supply his or her own estimates in the context of the particular statistical model being used. For this example, a

special code of (-99999) denotes missing observations. A score of -99999 is such an unlikely score in most social science applications that a mistake in the substitution process will be obvious.

The SPSS commands that perform this substitution process, inserting -99999 for scores on variables with missing data, are as follows:

```
SET BLANKS=-99999 PRINTBACK=LISTING .
TITLE 'MISSING DATA MEAN SUBSTITUTION EXAMPLE' .
DATA LIST FILE='a:missingdata.txt' RECORDS=1/
  ID 1-2 X1 to X6 4-9 .
```

The variable "missing" is then created for those entries of the data matrix now identified by the code -99999 as nonresponse items. That is, these commands substitute the score of -99999 for any numeric data fields that are blank. Appendix A presents the syntax file for this first phase.

The SPSS syntax commands that count the number of missing scores for each person, here using a new variable named "missing", are:

```
SUBTITLE 'Get n Missing for each person' .
COUNT MISSING=X1 to X6 (-99999) .
FREQUENCIES VARIABLES=MISSING .
LIST VARIABLES=ALL/CASES=9999/FORMAT=NUMBERED .
```

These commands also produce a summary of how many people have how many missing scores (i.e., the counts represented as scores on the variable named "missing"). The summary for the Table 1 data is presented in Table 2. In the present example,

as can be seen in Table 2, 8 people had 0 missing data, while 4 people had 1 piece of missing data.

INSERT TABLE 2 ABOUT HERE.

In a real data set there would be a wider range of counts of missing data, ranging conceivably from 0 to the number of data points in the data set. The results of the "frequency" analysis would be consulted at this point, prior to conducting any further analyses. For example, let's say that in a data set for 150 people and involving 75 items, the missing score counts were:

Missing	Frequency
0	126
1	11
2	8
3	3
15	1
27	1
Total	150

In such a data set the researcher must first decide how much missing data is permissible. This is a matter of thoughtful judgment. Perhaps estimating up to 3 scores for each person is reasonable. However, estimating 27 of 75 scores for the one person missing 27 scores is probably unreasonable.

Assume that the researcher decided to omit any case with more than 3 missing scores. The researcher would add the following SPSS syntax command to the syntax file:

```
SELECT IF (MISSING LT 4) .
```

In all subsequent analyses, after the execution of this command, any person with more than 3 missing scores will be omitted from any requested analyses.

Step Two: Find Means Using Selected Cases. In each phase of the procedure, new SPSS syntax commands are simply added on to the end of the existing SPSS syntax file. In this phase of the analysis a cutoff for an acceptable number of missing scores is declared, these cases are selected, and means on non-missing scores of only the subset of selected cases are computed. Appendix B presents the extended syntax file incorporating this phase of analysis for the present heuristic example.

For the present example, cases missing more than 1 score might be deleted (resulting for the Table 1 data set in no cases being omitted from subsequent analyses). Then means would be found for the non-missing scores of the selected cases, using the SPSS syntax commands:

```
SELECT IF (MISSING LT 2) .
MISSING VALUES X1 to X6(-99999) .
SUBTITLE 'Find means for each variable based on the number
  Of valid cases' .
DESCRIPTIVES VARIABLES=ALL .
```

Table 3 presents the results of this analysis for the present example.

INSERT TABLE 3 ABOUT HERE.

Step Three: Replace Missing Scores with Variable Means. In order for the following "IF" statements to work, the "COMMENT" command must be inserted before the syntax statement "MISSING VALUES X1 TO X6 (-99999)" and the command "EXECUTE" must be inserted after this statement. A series of "IF" statements are then added replacing the cases with missing data with variable means. Appendix C presents the syntax file for this last phase.

```
COMMENT MISSING VALUES X1 to X6(-99999) .
EXECUTE .
SUBTITLE 'Find means for each variable based on the number
of valid cases' .
DESCRIPTIVES VARIABLES=ALL .

IF (X1 lt -1) X1=3.73 .
IF (X2 lt -1) X2=5.91 .
IF (X3 lt -1) X3=5.18 .
IF (X4 lt -1) X4=7.36 .
IF (X5 lt -1) X5=4.25 .
IF (X6 lt -1) X6=4.92 .
EXECUTE .
MISSING VALUES X1 to X6(-99999) .
DESCRIPTIVES VARIABLES=ALL .
```

Table 4 presents the descriptive statistics produced by variable mean substitution.

INSERT TABLE 4 ABOUT HERE.

Regression Estimation

Estimation by developing a regression equation to predict the criterion of a variable with missing data using valid cases, and

then applying the equation to the valid scores on other variables of persons missing scores on that given variable, can also be used. This estimation is more sophisticated, because it takes into account relationships among the variables.

Regression methods rely on the information contained in the non-missing values of variables to provide estimates of the missing values for the variable of interest. Each variable with a missing value is, in turn, treated as a criterion variable and is regressed onto all the other variables having observed values to predict the criterion variable. There are many variations to the regression method depending on how many predictors are used and how the missing values on predictors themselves are handled (Kaiser, 1990).

In the application of a linear regression model, the assumption is made that the relationship between the incomplete variable and the covariates (predictors) on which it is regressed is linear over the full range of values. Furthermore, the reason for a value being missing is assumed to be unrelated to the value of the predictors. One needs access to nothing more than means, standard deviations, and correlations in order to compute estimates via regression (Raymond & Roberts, 1987).

As Thompson (1992) noted,

> Conventional regression analysis employs two types of weights: an additive constant ("a") applied to

every case and a multiplicative constant ("b") applied to the predictor variable for each case. Thus, the weighting system takes the form of a regression equation:

$$Y <\text{----} \hat{Y} = a + b\ (X)$$

For example, it is known that the following system of weights works reasonably well to predict height at age 21 from height at age 2:

$$Y <\text{----} \hat{Y} = 0 + 2.0\ (X)$$

Thus, an individual that is 27" tall at age 2 is predicted to have a height of 54" (0 + 2.0 x 27 = 0 + 54 = 54) at age 21. (p.6)

This missing data estimation is more efficient because a greater amount of available information is used. The computational complexity of the regression approach to missing data substitution had made it impractical previously; however, the necessary computer programs (e.g., SPSS) are now widely available. Using regression-estimated scores for missing data does not attenuate the relationships among variables, because a regression-estimated score will not equal the mean, unless (a) the squared multiple R for the regression equation is exactly 0 or (b) the predictor variable scores for non-missing data for cases with missing data

all exactly equal the respective means of these predictor variables.

SPSS for Windows: Estimation by Regression

The same heuristic data set of 12 participants is used to make the discussion more concrete and accessible. The substitution process is done in two phases: (a)listwise deletion produces an initial correlation matrix and each variable with missing values is in turn treated as a dependent variable and regressed on the non-missing variables, and then (b) the resulting regression equations are used to predict and replace the missing values.

Step One: Obtaining Regression Equations to Predict Missing Values. Appendix D lists the SPSS commands for obtaining regression equations to predict the missing values. The SPSS output to predict missing values on each variable are presented in Tables 5-8, respectively.

INSERT TABLES 5-8 ABOUT HERE.

Step Two: Using Output from First Phase to Estimate Missing Values. A series of "IF" statements are then added to the syntax file that plug in the multiplicative and additive weights for each variable. The syntax file is then rerun, producing estimates for the missing values for each variable and these estimates are inserted into the incomplete data set simultaneously. This

process is illustrated in Appendix E. The new data set resulting from regression estimation is presented in Table 9.

INSERT TABLE 9 ABOUT HERE.

Conclusion

Missing data are a practical problem and a practical solution is needed. Data often can be treated in some fashion so as to minimize information loss or sources of bias. Which technique is best depends on several factors.

This paper has presented four options for dealing with missing data. Listwise deletion and pairwise deletion methods both result in a reduction in sample size which leads to reduced precision in the estimates of the population parameters. This reduction in sample size also reduces the power of statistical significance testing, and this poses a potential threat to statistical conclusion validity (Orme & Reis, 1991). Although the same attenuation of the correlation coefficicient occur, the methods of inserting means and using regression analyses are about equally effective under conditions of low multicollinearity (Raymond & Roberts, 1987). The most important advantages of these mean imputation methods are the retention of sample size and, consequently of statistical power in subsequent analyses.

The preceding discussion and demonstration has hopefully alerted researchers to the possible complications arising from missing data and to the fact that they may be using less than an optimal solution if they use packaged (e.g., SPSS) defaults to address the problem. Unfortunately, because of the numerous factors influencing the relative success of the competing techniques, no one method for handling the missing data problem has been shown to be uniformly superior. Hence, as Anderson, Basilevsky, and Hum (1983) emphasized in their review of missing data procedures, "we return to the old precept that still holds true: The only real cure for missing data is to not have any" (p.480).

References

Anderson, A.B., Basilevsky, A. & Hum, D. P. J. (1983). Missing data: A review of the literature. In P. H. Rossi, J. D. Wright, & A. B. Anderson (Eds.), Handbook of survey research (pp. 415-494). San Diego: Academic Press.

Cohen, J., & Cohen, P. (1983). Missing data. In J. Cohen & P. Cohen, Applied multiple regression: Correlation analysis for the behavioral sciences (pp. 275-300). Hillsdale, NJ: Erlbaum.

Glasser, M. (1964). Linear regression analysis with missing observations among the independent variables. Journal of the American Statistical Association, 59, 834-844.

Kaiser, J. (1990, August). The robustness of regression and Substitution by mean methods in handling missing values. Paper presented at the second Islamic Countries Conference on Statistical Science, Malaysia.

Kim, J., & Curry, J. (1977). The treatment of missing data in multivariate analysis. Sociological Methods and Research, 6, 215-240.

Little, R.J.A. & Rubin, D.R. (1987). Statistical analysis with missing data. New York: Wiley.

Orme, J.G., & Reis, J. (1991). Multiple regression with missing data. Journal of Social Service Research, 15, 61-91.

Raymond, M.R., & Roberts, D.M. (1987). A comparison of methods for treating incomplete data in selection research.

Educational and Psychological Measurement, 47, 13-26.

Thompson, B. (1992, April). Interpreting regression results: beta weights and structure coefficients are both important. Paper presented at the annual meeting of the American Educational Research Association, San Francisco. (ERIC Document Reproduction Service No. ED 344 897)

Walsh, B.D. (1996). A note on factors that attenuate the correlation coefficient and its analogs. In B. Thompson (Ed.), Advances in social science methodology (Vol. 4, pp.21-32). Greenwich, CT: JAI Press.

Wilkinson, L., & The APA Task Force on Statistical Inference. (1999). Statistical methods in psychology journals: Guidelines and explanations. American Psychologist, 54, 594-604. [reprint available through the APA Home Page: http://www.apa.org/journals/amp/amp548594.html]

Witta, L. & Kaiser, J. (1991, November). Four methods of Handling missing data with the 1984 General Social Survey. Paper presented at the annual meeting of the Mid-South Educational Research Association, Lexington, KY.

Table 1

The Data Set as Would Appear in Word
File: "a:\missingdata.txt"

```
01  24890
02 2 6903
03 35 897
04 195 26
05 947605
06 386590
07 386809
08 939640
09 288957
10 120975
11 376508
12 590869
```

Table 2

SPSS Frequencies Table for Count of Missing Data Using the Variable Named "Missing"

Valid	Frequency	Percent	Valid Percent	Cumulative Percent
.00	8	66.7	66.7	66.7
1.00	4	33.3	33.3	100.0
Total	12	100.0	100.0	

Table 3

SPSS Descriptive Statistics Table Utilized to Find Variable Means

	Descriptive Statistics				
	N	Minimum	Maximum	Mean	Std. Deviation
ID	12	1	12	6.50	3.61
X1	11	1	9	3.73	2.83
X2	11	2	9	5.91	2.77
X3	11	0	9	5.18	2.89
X4	11	5	9	7.36	1.57
X5	12	0	9	4.25	3.77
X6	12	0	9	4.92	3.42
MISSING	12	.00	1.00	.3333	.4924
Valid N (listwise)	8				

Table 4

SPSS Final Descriptive Statistics Table with Substituted Variable Means

	Descriptive Statistics				
	N	Minimum	Maximum	Mean	Std. Deviation
ID	12	1	12	6.50	3.61
X1	12	1	9	3.73	2.70
X2	12	2	9	5.91	2.64
X3	12	0	9	5.18	2.76
X4	12	5	9	7.36	1.49
X5	12	0	9	4.25	3.77
X6	12	0	9	4.92	3.42
MISSING	12	.00	1.00	.3333	.4924
Valid N (listwise)	12				

Note. After the mean substitutions, the n's for variables X1 through X4 are all now 12, rather than the n=11 reported in Table 3. Also, because means have been substituted, the means for the variables remain the same as they were in Table 3 (e.g., 3.73 and 3.73 for variable X1 in Tables 3 and 4, respectively).

However, mean substitution makes the scores less spread out, because scores are added toward the middle of the distribution. For example, in Table 3 the SD of X1 is 2.83, but after the mean substitution in this table the SD of X1 is reported to be 2.70. Of course, these changes would be less dramatic with large data sets, or data sets with proportionately fewer missing scores.

Table 5

SPSS Output to Predict Missing Values on X1

Variable	B	SE B	Beta	T	Sig T
x2	.524817	1.414864	.461211	.371	.7463
x3	-.535093	1.156447	-.595180	-.463	.6890
x4	.663248	2.051605	.365452	.323	.772
x5	-1.246413	1.629975	-1.431360	-.765	.5244
x6	-1.377397	1.991655	-1.644273	-.692	.5607
(constant)	11.560359	9.764212		1.184	.3581

Table 6

SPSS Output to Predict Missing Values on X2

Variable	B	SE B	Beta	T	Sig T
X1	.122646	.330644	.139560	.371	.7463
X3	.690676	.327832	.874180	2.107	.1697
X4	-1.168327	.593751	-.732533	-1.968	.1880
X5	1.104418	.438791	1.443207	2.517	.1282
X6	1.377414	.447256	1.871058	3.080	.0912
(Constant)	-1.542556	6.058558		-.255	.8228

Table 7

SPSS Output to Predict Missing Values on X3

Variable	B	SE B	Beta	T	Sig T
X1	-.180710	.390551	-.162466	-.463	.6890
X2	.998114	.473759	.788594	2.107	.1697
X4	1.216489	.869387	.602621	1.399	.2967
X5	-1.367642	.473665	-1.412020	-2.887	.1019
X6	-1.632382	.572401	-1.751934	-2.852	.1041
(Constant)	5.485392	6.302347		.870	.4759

Table 8

SPSS Output to Predict Missing Values on X4

Variable	B	SE B	Beta	T	Sig T
X1	.074875	.231609	.135889	.323	.7772
X2	-.564390	.286827	-.900154	-1.968	.1880
X3	.406647	.290618	.820883	1.399	.2967
X5	.726869	.351360	1.514917	2.069	.1745
X6	.908178	.377495	1.967573	2.406	.1379
(Constant)	.296341	4.273484		.069	.9510

Table 9

SPSS Output Using Regression Equations to Estimate Missing Values

ID	X1	X2	X3	X4	X5	X6	MISSING
1	4.55789	2.00000	4.00000	8.00000	9	0	1.00
2	2.00000	-3.53591	6.00000	9.00000	0	3	1.00
3	3.00000	5.00000	-4.06971	8.00000	9	7	1.00
4	1.00000	9.00000	5.00000	4.22775	2	6	1.00
5	9.00000	4.00000	7.00000	6.00000	0	5	.00
6	3.00000	8.00000	6.00000	5.00000	9	0	.00
7	3.00000	8.00000	6.00000	8.00000	0	9	.00
8	9.00000	3.00000	9.00000	6.00000	4	0	.00
9	2.00000	8.00000	8.00000	9.00000	5	7	.00
10	1.00000	2.00000	.00000	9.00000	7	5	.00
11	3.00000	7.00000	6.00000	5.00000	0	8	.00
12	5.00000	9.00000	.00000	8.00000	6	9	.00

Appendix A

Using the COUNT Command to Get Number of Missing Values

```
set blanks=-99999 printback=listing .
title 'MISSING DATA MEAN SUBSTITUTION EXAMPLE' .
data list file='a:missingdata.txt' records=1/
  ID 1-2 X1 to X6 4-9 .
subtitle 'GET n MISSING FOR EACH PERSON' .
count missing=X1 to X6 (-99999) .
frequencies variables=missing .
list variables=all/cases=9999/format=numbered .
```

Appendix B

Deletion of Some Cases and Determining Means

```
set blanks=-99999 printback=listing .
title 'MISSING DATA MEAN SUBSTITUTION EXAMPLE' .
data list file='a:missingdata.txt' records=1/
  ID 1-2 X1 to X6 4-9 .
subtitle 'GET n MISSING FOR EACH PERSON' .
count missing=X1 to X6 (-99999) .
frequencies variables=missing .
list variables=all/cases=9999/format=numbered .

select if (missing lt 2) .
missing values X1 to X6(-99999) .
subtitle 'Find means for each variable based on the number of
valid cases' .
descriptives variables=all .
```

Appendix C

Perform the Mean Substitution

```
set blanks=-99999 printback=listing .
title 'MISSING DATA MEAN SUBSTITUTION EXAMPLE' .
data list file='a:missingdata.txt' records=1/
  ID 1-2 X1 to X6 4-9 .
subtitle 'GET n MISSING FOR EACH PERSON' .
count missing=X1 to X6 (-99999) .
frequencies variables=missing .
list variables=all/cases=9999/format=numbered .

select if (missing lt 2) .
comment missing values X1 to X6(-99999) .
execute .
subtitle 'Find means for each variable based on the number of
valid cases' .
descriptives variables=all .

if (X1 lt -1) X1=3.73 .
if (X2 lt -1) X2=5.91 .
if (X3 lt -1) X3=5.18 .
if (X4 lt -1) X4=7.36 .
if (X5 lt -1) X5=4.25 .
if (X6 lt -1) X6=4.92 .
execute .
missing values X1 to X6(-99999) .
descriptives variables-all .
```

Appendix D

Obtaining regression equations to predict missing values

```
Set blanks=-99999 UNDEFINED=WARN printback=listing.
title 'MISSING DATA REGRESSION EXAMPLE' .
Data list file='a:missingdata.txt' fixed records=1 table
  /1 ID 1-2 X1 to X6 4-9 .
subtitle 'GET n MISSING FOR EACH PERSON' .
count missing=X1 to X6(-99999) .
frequency variables=missing .
list variables=all/cases=9999/format=numbered .
subtitle '1a  Find equation to predict missing on X1' .
temporary .
select if (missing eq 0) .
regression variables=X1 to X6/dependent=x1/
  enter X2 to X6 .
subtitle '2a  Find equation to predict missing on X2' .
temporary .
select if (missing eq 0) .
regression variables=X1 to X6/dependent=X2/
  enter X1 X3 to X6 .
subtitle '3a  Find equation to predict missing on X3' .
temporary .
select if (missing eq 0) .
regression variables=X1 to X6/dependent=X3/
  enter X1 X2 X4 to X6 .
subtitle '4a  Find equation to predict missing on X4' .
temporary .
select if (missing eq 0) .
regression variables=X1 to X6/dependent=X4/
  enter X1 to X3 X5 X6 .
```

Appendix E

Using output from first phase of analysis to find missing values

```
Set blanks=-99999 UNDEFINED=WARN printback=listing.
title 'MISSING DATA REGRESSION EXAMPLE' .
Data list file='a:missingdata.txt' fixed records=1 table
  /1 ID 1-2 X1 to X6 4-9 .
subtitle 'GET n MISSING FOR EACH PERSON' .
count missing=X1 to X6(-99999) .
frequency variables=missing .
list variables=all/cases=9999/format=numbered .
subtitle '1a  Find equation to predict missing on X1' .
temporary .
select if (missing eq 0) .
regression variables=X1 to X6/dependent=X1/
  enter X2 to X6 .
subtitle '2a  Find equation to predict missing on X2' .
temporary .
select if (missing eq 0) .
regression variables=X1 to X6/dependent=X2/
  enter X1 X3 to X6 .
subtitle '3a  Find equation to predict missing on X3' .
temporary .
select if (missing eq 0) .
regression variables=X1 to X6/dependent=X3/
  enter X1 X2 X4 to X6 .
subtitle '4a  Find equation to predict missing on X4' .
temporary .
select if (missing eq 0) .
regression variables=X1 to X6/dependent=X4/
  enter X1 to X3 X5 X6 .

IF (X1 LT -1)X1=(.524817 * X2) + (-.535093 * X3) +
                (.663248 * X4) + (-1.246413 * X5) +
                (-1.377397 * X6) + 11.560359 .
IF (X2 LT -1)X2=(.122646 * X1) + (.690676 * X3) +
                (-1.168327 * X4) + (1.104418 * X5) +
                (1.377414 * X6) + -1.542556 .
IF (X3 LT -1)X3=(-.180710 * X1) + (.998114 * X2) +
                (1.216489 * X4) + (-1.367642 * X5) +
                (-1.632382 * X6) + 5.485392 .
IF (X4 LT -1)X4=(.074875 * X1) + (-.564390 * X2) +
                (.406647 * X3) + (.726869 * X5) +
                (.908178 * X6) + .296341 .
PRINT FORMATS X1 TO X4(F8.5) .
LIST VARIABLES=ALL/CASES=9999/FORMAT=NUMBERED .
```

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

University of Maryland
ERIC Clearinghouse on Assessment and Evaluation
1129 Shriver Laboratory
College Park, MD 20742
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

ERIC Processing and Reference Facility
1100 West Street, 2nd Floor
Laurel, Maryland 20707-3598

Telephone: 301-497-4080
Toll Free: 800-799-3742
FAX: 301-953-0263
e-mail: ericfac@inet.ed.gov
WWW: http://ericfac.piccard.csc.com

EFF-088 (Rev. 9/97)
PREVIOUS VERSIONS OF THIS FORM ARE OBSOLETE.